APENet: LQCD clusters à la APE

R. Ammendola[1] D. Rossetti[2]

[1]Istituto Nazionale di Fisica Nucleare, Sezione Roma II

[2]Istituto Nazionale di Fisica Nucleare, Sezione Roma I

Lattice 04, Chicago – 23rd June 2004

Collaboration

The work presented in this talk have been developed by

- Roberto Ammendola
- Marco Guagnelli
- Giuseppe Mazza
- Filippo Palombi
- Roberto Petronzio
- Davide Rossetti
- Andrea Salamon
- Piero Vicini

for INFN Roma 1 and Roma 2
The group is involved in hardware, software and application code development

Framework

Local and homogeneous algorithms → APE Supercomputers family

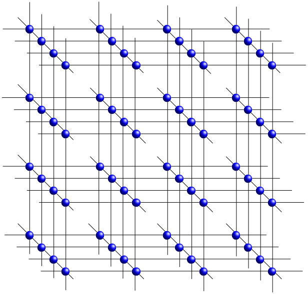
- 3D point-to-point interconnect
- custom processors

Clusters of commodity processors:

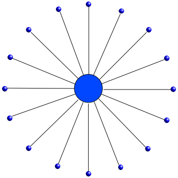
- improving number crunching
- networking lagging behind

→ Need for a custom interconnect system

Target: medium installations (64 nodes and over)



3D mesh topology



star topology

APENet

APENet is a 3D network of point-to-point links with toroidal topology.

- Each computing node has 6 bi-directional full-duplex communication channels
- Computing nodes are arranged in a 3D cubic mesh
- Data is transmitted in packets which are routed to the destination node
- No external router device is necessary
- Lightweight low level protocol
- Internal cut-through switching capabilities

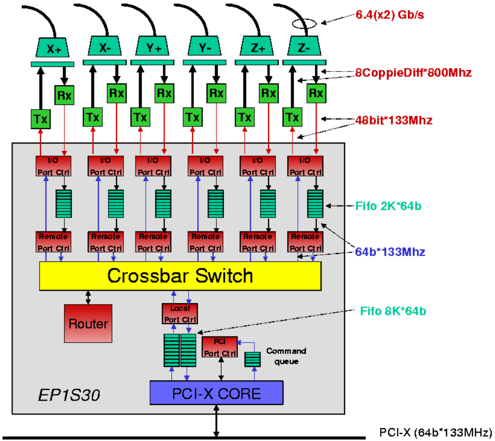
Architecture

The building block of APENet is the APELink Card
APELink is a PCI-X 133 MHz 64 bit card
Three major blocks:

- PCI-X interface
- Crossbar Switch
- Communication Links

The Crossbar Switch is divided in:

- a Switch
- an Arbiter
- a Router

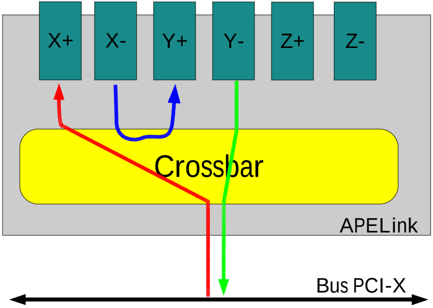


Hopping

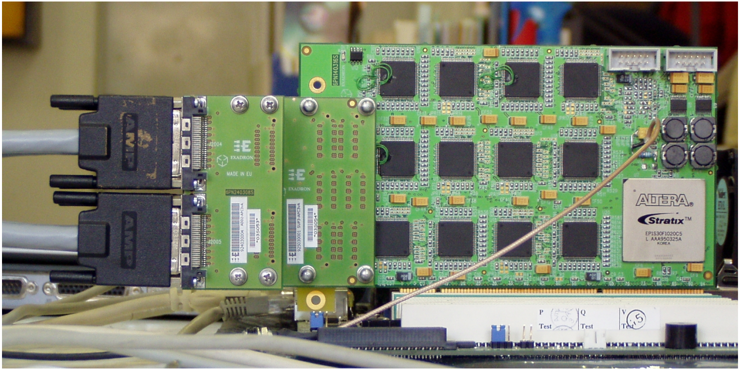
Multiple hop transactions are possible

- each switch can support up to three simultaneous connections
- minumum hardware latencies for multiple hop transmissions, order of 10 clock cycles per hop

e.g. `ape_sndrcv(X_PLUS, Y_MINUS)`



Hardware Components



- Altera Stratix EP1S30, 1020 pin package, fastest speed grade
- National Serializers/Deserializers DS90CR485/486, 48 bit 133 MHz

Usage of a programmable device allows possible logic redesign and quick on-field firmware upgrade.

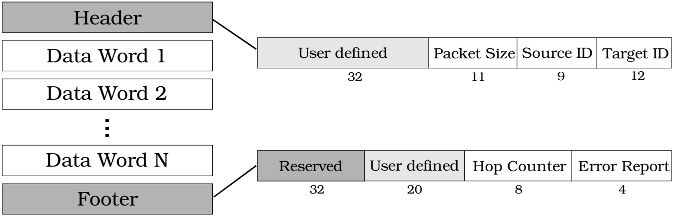
Packet and data encoding

$BER < 10^{-14} \rightarrow$ Link reliability allows highly efficient low level encoding scheme

- a simple parity checking code is used
- data payload is 90 % – signaling is 10 %

Lightweight transfer protocol:

- routing informations are stored in the header
- errors are reported in the footer



APENet Software

Software development:

- APELink kernel device driver
- low level C API
- mid level C API
- experimental network device driver
- LAM/MPI porting

Mid level API is targeted for numerical application code

- ape_send()
- ape_recv()
- ape_sndrcv()
- ape_broadcast()
- ape_global_sum()
- ...

Software is developed under GPL policy

Testbed

Preliminary benchmarks:

- software development in progress
- hardware in prototypal state

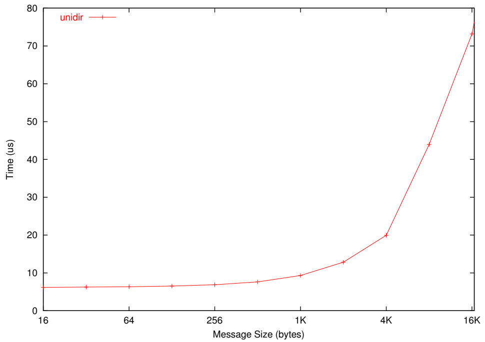
Test are performed on a 4 nodes cluster

- dual Xeon 2.4-3.0 GHz
- various chipsets
- Linux Kernel 2.4.latest
- PCI-X clock 100-133 MHz
- Link clock 100 MHz

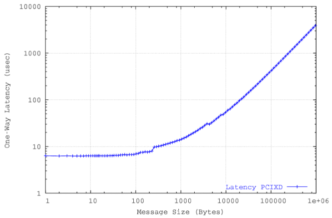
In these conditions asymptotic unidirectional bandwidth is 508 MB/sec
Two standard MPI-like micro-benchmarks have been ran using `ape_send()`, `ape_recv()` and `ape_sndrcv()` high level functions:

- Bandwidth
- Latency

Measured Latency

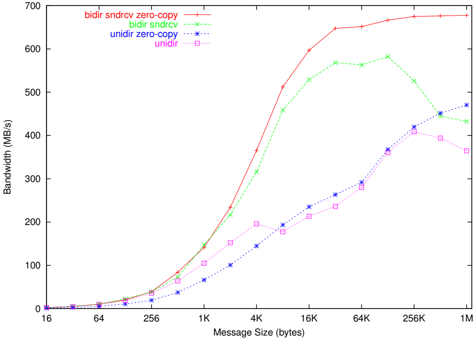


APENet Latency: 6.2 μsec

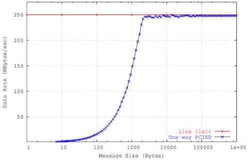


Myrinet Latency: 6.3 μsec

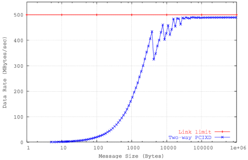
Measured Bandwidth



APENet Unidirectional Bandwidth: 470 *MB*/*sec*
APENet Bidirectional Bandwidth: 670 *MB*/*sec*



Myrinet Unidirectional Bandwidth:
248 *MB*/*sec*



Myrinet Bidirectional Bandwidth:
489 *MB*/*sec*

Conclusions and Outlooks

Summarizing APENet:

- 6 bi-directional links
- measured latency: 6.2 μsec
- measured uni-directional bandwidth: 470 MB/sec
- measured bi-directional bandwidth: 670 MB/sec

Perfomance measurements on physics code has already started

Increasing of performance is expected soon:

- Links fully working at 133 MHz
- Improving the low level device driver
- Implementing the network driver

A 16 nodes cluster will be set up in the following weeks
An upgrade to 64 nodes is programmed by the end of the year